ABSTRACT

THESIS: A Two-Step Integrated Approach to Detect Differentially Expressed Genes in RNA-Seq Data

STUDENT: Naim Al Mahi

DEGREE: Master of Science

COLLEGE: Sciences and Humanities

DATE: May, 2014

PAGES: 52

Motivation: RNA-Sequence or RNA-Seq experiments produce millions of discrete DNA sequence reads, as a measure of gene expression levels. It enable researchers to investigate complex aspects of the genomic studies. These include but not limited to identification of differentially expressed (DE) genes in two or more treatment conditions and detection of novel transcripts. One of the common assumptions of RNA-Seq data is that, all gene counts follow an overdispersed Poisson or negative binomial (NB) distribution which is sometimes misleading because some genes may have stable transcription levels with no overdispersion. Thus, a more realistic assumption in RNA-Seq data is to consider two sets of genes: overdispersed and non-overdispersed.

Method: We propose a new two step integrated approach to detect differentially expressed (DE) genes in RNA-Seq data using standard Poisson model for non-overdispersed genes and NB model for overdispersed genes. This is an integrated approach because this method can be combined with any other NB based methods for detecting DE genes.

Results: We evaluate the proposed approach using two simulated and two real RNA-Seq data sets. We compare the performance of our proposed method combined with the four popular R-software packages edgeR, DESeq, sSeq, and NBPSeq with their default settings. For both the simulated and real data sets, integrated approaches perform better or at least equally well compared to the regular methods embedded in these R-packages.